The Corporate Project Economy:

Large-Scale Extraction and Econometric Analysis of 19,000 Strategic Initiatives from Two Decades of SEC Filings (2010–2026)

Manel Kammoun
manel.kammoun@uqo.ca

Djerrycharli Tandjambianda
djerrycharli.tandjambianda@uqo.ca

Simon-Pierre Boucher
simon-pierre.boucher@uqo.ca
Département des sciences administratives
Université du Québec en Outaouais

Gatineau – Pavillon Alexandre-Taché
283, boulevard Alexandre-Taché
Gatineau, Québec, Canada J9A 1L8

June 26, 2026

Abstract

Corporate strategy is enacted through *projects* — plant builds, data centers, acquisitions, research programs, digital migrations — yet these initiatives are scattered, unstructured, and buried inside the narrative sections of regulatory filings. We introduce SPID (the SEC Project Intelligence Database), a pipeline that converts the raw text of corporate filings into a structured, project-centric panel, and we use it to build the largest machine-extracted census of corporate projects we are aware of: **19,227 resolved projects** drawn from **39,930 mentions** across **153,894** filing sections (10-K, 10-Q and 8-K) filed by **500 large U.S. public companies** between 2010 and 2026. A large-language-model extractor (`gpt-4o-mini`, JSON mode) maps each section to a controlled 21-category project taxonomy with quantitative attributes (announced capital, status, locations, technologies, partners), which are then resolved into firm-level project portfolios with timelines, a 32,226-node knowledge graph, and 384-dimensional semantic embeddings. Using this panel we estimate a battery of models: fixed-effects Poisson and negative-binomial count models of project-announcement intensity; hedonic log-capital regressions with firm and year effects; a logistic technology-diffusion model with an estimated inflection in **2023**; a difference-in-differences event study of COVID-19 and corporate digitalization; Cox proportional-hazards models of project lifecycles; community detection over a technology co-occurrence network; unsupervised clustering of project embeddings; and a gradient-boosted

classifier of "mega-projects" that attains a cross-validated AUC of **0.74** from text-derived features alone. We document strong persistence in project activity, a sharp post-2020 acceleration of announcements, an energy-transition technology cluster at the center of the corporate-innovation network, and project lifecycles that range from a median of four years for M&A integration to twelve years for digital and AI programs. Beyond the substantive findings, the paper demonstrates that LLM-based information extraction can turn the unstructured regulatory record into research-grade panel data for empirical finance and the economics of innovation.

Keywords: corporate investment, text-as-data, large language models, information extraction, innovation, SEC filings, knowledge graphs, panel econometrics.
JEL: G31, O30, C55, D22, M15.

Contents

1 Introduction **4**

2 Related Literature **6**

3 Data and Extraction Pipeline **7**

3.1 Source corpus 7

3.2 LLM extraction 7

3.3 Resolution, graph, and embeddings 7

3.4 The economics of extraction 8

3.5 Analysis datasets 8

4 Descriptive Evidence and Stylized Facts **9**

4.1 Six stylized facts 13

5 Econometric Framework **13**

6 The Extensive Margin: Count Models **15**

7 The Intensive Margin: What Drives Announced Capital? **16**

8 Technology Diffusion: The Rise of Corporate AI **18**

9 COVID-19 and Digitalization: A Difference-in-Differences Test **19**

10 Project Lifecycles: Survival Analysis **20**

11 The Innovation Network **22**

12 Latent Themes: Embedding the Project Corpus **24**

13 Predicting Mega-Projects **25**

14 Robustness and Validation **26**

15 Limitations **27**

16 Discussion and Implications **28**

17 Conclusion **29**

A Project Taxonomy and Variable Definitions **30**

B Pipeline and Reproducibility **30**

1 Introduction

The modern public corporation is, in operational terms, a portfolio of *projects*. New products are launched, plants and data centers are built, firms are acquired and integrated, enterprise systems are migrated to the cloud, research pipelines are advanced, and entire business lines are reoriented around new technologies such as artificial intelligence. These initiatives are the mechanism through which strategy becomes capital expenditure and, eventually, realized returns. Yet for the empirical researcher they are remarkably hard to observe. Financial statements report aggregate capital expenditure, not the projects that compose it; press releases are selective and unstructured; and the richest single source — the narrative discussion inside regulatory filings — is written in prose, not in tables.

This paper has two goals. The first is methodological: to show that recent advances in large language models (LLMs) make it possible to convert the unstructured regulatory record into structured, research-grade panel data at scale and at low cost. The second is substantive: to use the resulting data to characterize the *corporate project economy* of large U.S. firms over the 2010–2026 period and to estimate a series of econometric models that speak to classic questions in corporate investment and the economics of innovation.

We construct SPID, the SEC Project Intelligence Database, on top of a 500-company corpus of filings downloaded from EDGAR. The pipeline (i) sections raw 10-K HTML into its project-dense items (Business, Risk Factors, Properties, MD&A), (ii) sends every project-relevant text window to an LLM extractor that returns a strict-JSON project record against a 21-category taxonomy, (iii) resolves the resulting mentions into firm-level projects with timelines using stable hashed identifiers, and (iv) builds a knowledge graph and a set of semantic embeddings on top of the resolved projects. The end product is a panel of **19,227 projects** and **39,930 dated mentions** spanning **500 firms**, **11 GICS sectors**, **21 project types**, and **3,080 distinct locations**, of which **9,321 (48.5%)** carry an explicit dollar amount with a median announced value of **US$718 million**.

The scale and structure of this panel allow us to ask questions that aggregate data cannot answer. *How persistent is project activity within firms, and did it change after 2020?* We estimate fixed-effects Poisson and negative-binomial models and find strong, statistically robust persistence and a significant post-COVID jump in announcement intensity, against a mild negative secular time trend once firm composition is held fixed. *What kinds of projects command the most capital?* Hedonic log-capital regressions with firm and year fixed effects explain roughly a third of the variation in announced amounts and reveal large premia for capital-intensive project families relative to M&A integration. *How fast is corporate AI adoption diffusing?* A logistic diffusion model places the inflection point of cumulative AI adoption in **2023**, coincident with the generative-AI boom, and the AI share of new project announcements rises from 0.7% in 2015 to 4.7% in 2025. *Did the pandemic differentially accelerate the digitalization of physically intensive incumbents?* A difference-in-differences event study with clean parallel pre-trends returns a precisely estimated null, cautioning against the popular narrative of pandemic-driven digital convergence. *How long do projects live?* Cox and Kaplan–Meier models show that M&A integrations resolve fastest (median

four years) while digital and AI programs persist longest (median twelve years), and that larger projects complete more slowly.

We complement these regressions with network and representation-learning analyses. A technology co-occurrence network exhibits strong community structure (modularity 0.52) organized into four "technology ecosystems," with energy-transition technologies — solar, wind, battery storage — occupying the most central positions in the corporate-innovation graph. Unsupervised clustering of the 384-dimensional project embeddings recovers six interpretable themes that map cleanly onto the taxonomy while crossing sector boundaries. Finally, a gradient-boosted classifier predicts whether a project is a "mega-project" (above the $718m median) with a cross-validated AUC of 0.74 using only text-derived structural features, with project type, announcement year and sector the most informative predictors.

Contributions. We make four contributions. First, we provide a *reproducible, low-cost methodology* for turning the narrative sections of regulatory filings into structured project-level panel data using an LLM extractor with a controlled taxonomy and deterministic resolution. Second, we deliver a *novel dataset* — to our knowledge the largest machine-extracted corporate-project census — together with a knowledge graph and embeddings that support search, analytics, and downstream modeling. Third, we present a *comprehensive econometric analysis* that combines count models, hedonic regressions, diffusion curves, difference-in-differences, survival analysis, network science, and machine learning on a single coherent panel. Fourth, we surface *substantive findings* about persistence, the post-2020 surge, AI diffusion timing, project longevity, and the centrality of the energy transition that are difficult or impossible to obtain from conventional structured data.

Caveats up front. Our census is an extraction from disclosure, not a ground-truth registry. Announced amounts are self-reported and heterogeneous in scope; the headline aggregate of announced capital should be read as an *upper-bound order of magnitude* rather than a netted investment total, because the same multi-year program can be mentioned (and counted) in several filings and because large M&A enterprise values dominate the sum. Throughout, we therefore emphasize *within-firm* and *relative* comparisons, which are robust to these level issues, and we treat the dollar figures as ordinal signals of project scale. Section 15 discusses measurement, selection, and extraction error in detail.

The remainder of the paper proceeds as follows. Section 2 situates the work in the text-as-data and corporate investment literatures. Section 3 describes the corpus, the extraction pipeline, and the resolution procedure. Section 4 presents descriptive statistics and stylized facts. Section 5 lays out the econometric framework. Sections 6–13 report results for each family of models. Section 14 collects robustness checks, and Section 15 discusses limitations. Section 17 concludes.

2 Related Literature

Our work sits at the intersection of three literatures: text-as-data methods in economics and finance, the empirical analysis of corporate investment, and the measurement of technological change and innovation.

Text as data. A now-large body of work extracts economic signals from corporate disclosure. Early contributions used dictionary methods and readability indices to study tone and sentiment in 10-K filings (Loughran & McDonald, 2011; Li, 2008), while later work measured risk and attention from the co-occurrence of words and topics (Hoberg & Phillips, 2016; Hassan et al., 2019). Gentzkow et al. (2019) survey the econometrics of text and emphasize the move from counting words to learned representations. We extend this trajectory in two ways. First, rather than scoring documents, we perform *structured information extraction*: each text window is mapped to a typed record with named entities and quantities. Second, we use a generative LLM as the extraction engine, which removes much of the brittle feature engineering required by earlier rule-based systems while retaining a controlled output schema.

LLMs for extraction and measurement. Recent papers show that LLMs can classify and extract economic content with accuracy approaching that of trained human readers, at a fraction of the cost (Korinek, 2023; Bybee et al., 2023). The methodological question is no longer whether extraction is feasible but how to make it *reproducible, auditable, and cheap at corpus scale*. Our pipeline addresses this with a deterministic taxonomy, JSON-mode decoding, confidence thresholds, and stable hashed mention identifiers that make the entire process incremental and resumable. The marginal cost of extracting a project mention is on the order of a tenth of a U.S. cent, which is what makes a census of this size practical.

Corporate investment. The empirical investment literature has long worked with aggregate capital expenditure and Tobin's q (Hayashi, 1982; Fazzari et al., 1988), and more recently with the rising importance of intangible and innovation-related investment (Peters & Taylor, 2017; Crouzet & Eberly, 2019). A persistent difficulty is that the unit of decision — the project — is not observed. Our data make the project the unit of analysis, allowing us to study the *extensive margin* (how many initiatives a firm announces) separately from the *intensive margin* (how large each is), and to characterize the *composition* of investment across project types and technologies.

Innovation and diffusion. The economics of technology diffusion has modeled adoption with logistic and Bass-type curves (Griliches, 1957; Bass, 1969), and a recent literature measures the diffusion of artificial intelligence across firms using job postings, patents, and earnings calls (Babina et al., 2024; Acemoglu et al., 2022). We contribute a disclosure-based measure of AI *projects* and estimate its diffusion path directly, finding an inflection in 2023 that aligns with the arrival of large generative models.

Networks and representations. Finally, we connect to work that represents firms and technologies as networks (Hidalgo et al., 2007) and as learned embeddings. Our knowledge graph and sentence-embedding clustering let us recover technology ecosystems and latent project themes that cut across the formal taxonomy, complementing the regression evidence with structure that is not imposed a priori.

3 Data and Extraction Pipeline

3.1 Source corpus

The underlying corpus is a 500-company subset of U.S. SEC filings retrieved from EDGAR, comprising raw filing HTML and a pre-parsed store of filing metadata and sectioned text. Two annual and quarterly forms (10-Q and 8-K) were already sectioned upstream; the project-dense 10-K filings were retained as raw HTML. As the first stage of SPID we ingested and sectioned **7,654** 10-K filings, extracting their Business (Item 1), Risk Factors (Item 1A), Properties (Item 2), and Management's Discussion and Analysis (Item 7/7A) sections into **27,164** clean text sections. Unioned with the pre-parsed 10-Q and 8-K sections, the analysis feed contains **153,894** filing sections (with at least 30 words) eligible for extraction: 103,926 from 8-K, 27,164 from 10-K, and 22,804 from 10-Q filings.

3.2 LLM extraction

Each eligible section is passed to an extraction model (`gpt-4o-mini`, temperature 0, JSON mode). To bound cost and focus the model, we first apply a deterministic "project gate" — a regular-expression filter for project-signal vocabulary — and send only the windows of text around matches, capped at roughly 14,000 characters per call. The model is instructed to return a strict-JSON array of project objects, each with a `project_name`, a `project_type` drawn from a fixed 21-category taxonomy (Table 2), a one-sentence grounded `description`, an `objective`, a `status` (planned / in-progress / completed / mentioned), an optional `amount_usd`, and arrays of `locations`, `technologies`, `partners`, `benefits`, `risks`, and `years`, together with a self-reported `confidence` in $[0, 1]$. The system prompt forbids speculation and instructs the model to prefer a small number of high-quality records over many vague ones; records below a confidence of 0.45 are discarded.

This stage produced **39,930 project mentions**. Because the model emits typed quantities, downstream analysis requires no further parsing: amounts are numeric, statuses and types are categorical, and list fields are stored as native arrays in a columnar database.

3.3 Resolution, graph, and embeddings

The same physical project is typically mentioned in several filings across years. We resolve mentions into projects by grouping on firm, project type, and a normalized project anchor, and we assign each project a timeline, a canonical location, an aggregated announced amount, first/last-seen dates, and

an average extraction confidence. Resolution collapses the 39,930 mentions into **19,227 distinct projects**; the mention timeline is retained for the survival analysis. From the resolved projects we build a directed knowledge graph with **32,226 nodes** (19,227 projects, 7,640 partners, 3,080 locations, 1,779 technologies, 500 companies) and **33,176 edges** (`owns`, `uses`, `located_in`), and we encode every project's text into a 384-dimensional sentence embedding (MiniLM, `all-MiniLM-L6-v2`) for semantic search and clustering.

3.4 The economics of extraction

A defining feature of the approach is its cost structure. Each section incurs at most one model call, and the deterministic project gate suppresses calls on sections with no project vocabulary; the focused-window construction caps the prompt at roughly 14,000 characters regardless of section length. The marginal cost of processing a section is therefore on the order of a tenth of a U.S. cent, and the full-corpus pass over 153,894 sections completed in a few hours on commodity hardware with bounded memory, batched into 34 firm-cohorts with checkpointing so that the run is fully resumable. This matters substantively: the feasibility of a *census* (as opposed to a sample) is precisely what distinguishes LLM extraction from earlier hand-coding or rule-based approaches, and it is what allows us to populate a 7,370-observation firm-year panel with project-level granularity. The same OpenAI-compatible interface can target a self-hosted open-weights model, so the pipeline is not tied to any single provider, and the per-section cost can in principle be driven to near zero on local infrastructure.

3.5 Analysis datasets

For econometric analysis we derive three datasets from the resolved store: (i) a project-level cross-section (19,227 rows) with engineered flags for AI and digital orientation, project age, and technology counts; (ii) a mention-level event table (39,930 rows) with filing dates and forms; and (iii) a balanced **firm-year panel** of **7,370** observations spanning 500 firms and the years 2010–2025 (the partial year 2026 is excluded from the panel), in which each firm-year records the count of newly announced projects, the number that are AI- or digital-oriented, announced capital, and project-type composition. Announcement year is taken from the project's first observed year. Table 1 reports summary statistics for the main variables.

Table 1: Summary statistics. Amounts are in millions of USD; duration is the span between a project's first and last appearance in filings.

Variable	N	Mean	SD	Median	Min	Max
Projects per firm	500	38.45	24.46	34.00	1.00	169.00
Mentions per project	19,227	2.08	3.95	1.00	1.00	185.00
Filings per project	19,227	1.88	2.95	1.00	1.00	185.00
Technologies per project	19,227	0.26	0.82	0.00	0.00	16.00
Announced amount (USD m)	9,321	2,520.55	11,794.72	718.00	0.00	898,200.00
Project duration (days)	5,905	1,401.60	1,479.37	734.00	1.00	5,957.00
Avg. confidence	19,227	0.89	0.10	0.90	0.50	1.00
Projects/firm-year (panel)	7,370	2.45	2.61	2.00	0.00	21.00
AI projects/firm-year	7,370	0.04	0.26	0.00	0.00	6.00

4 Descriptive Evidence and Stylized Facts

We begin with the broad contours of the corporate project economy.

Volume and trend. Figure 1 plots the number of extracted project mentions per year by filing form. The series is large and stable — between roughly 2,000 and 2,800 mentions per year — with a mild upward drift (a compound annual growth rate of about 2.2% over 2010–2025) and a visible step-up after 2020. 8-K filings (event-driven disclosures) and 10-K filings (annual reports) are the dominant sources, consistent with the fact that new initiatives are announced in 8-Ks and reviewed comprehensively in 10-Ks.

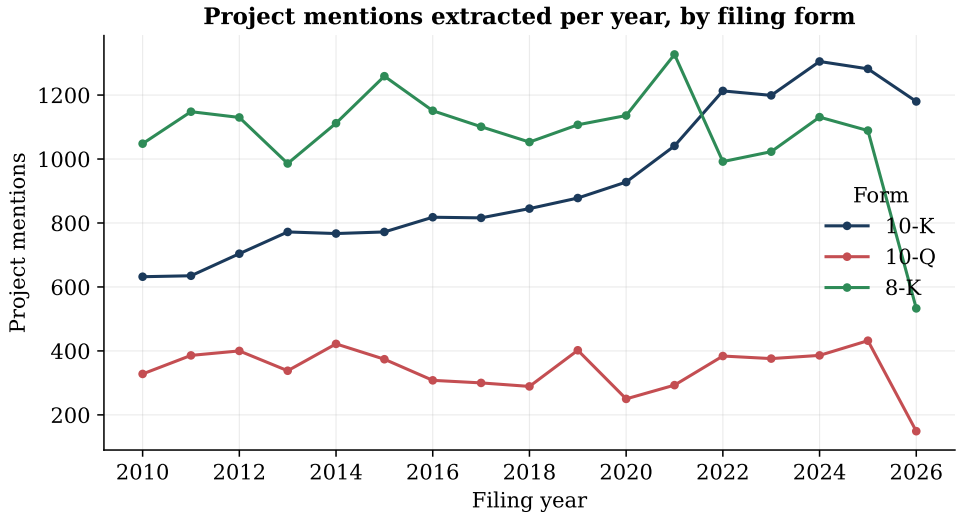


Figure 1: Project mentions extracted per year, by filing form. The corpus is roughly balanced across years, with 8-K and 10-K filings supplying the bulk of project signals.

Composition. Figure 2 and Table 2 show how projects distribute across the 21-category taxonomy. M&A integration is by far the most frequent category (7,274 projects), followed by capital-

expenditure programs (2,085), plant construction (1,252), and partnerships (1,239). The long tail includes specialized categories such as drug pipelines, ERP implementations, and data-center builds. The dominance of M&A integration reflects both the economic importance of acquisitions among large firms and the prominence of integration narratives in their filings.

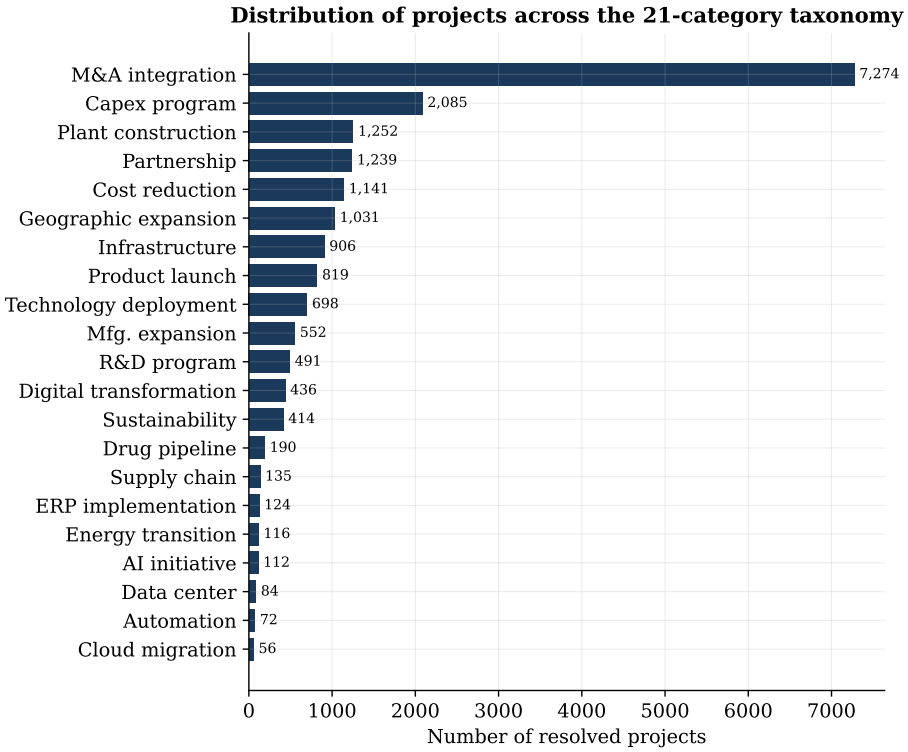


Figure 2: Distribution of resolved projects across the 21-category taxonomy.

Table 2: Projects by type: counts, aggregate announced capital, AI share, and mean extraction confidence. Capital is the sum of disclosed amounts and is an upper bound (see Section 15).

Project type	Projects	Capital ($bn)	AI share	Mean conf.
M&A integration	7,274	14,864.9	0.01	0.93
Capex program	2,085	4,602.6	0.00	0.93
Plant construction	1,252	463.0	0.01	0.87
Partnership	1,239	440.4	0.02	0.85
Cost reduction	1,141	725.7	0.00	0.86
Geographic expansion	1,031	175.0	0.00	0.82
Infrastructure	906	1,002.3	0.00	0.84
Product launch	819	48.8	0.03	0.86
Technology deployment	698	190.4	0.05	0.83
Mfg. expansion	552	145.6	0.01	0.84
R&D program	491	235.3	0.04	0.82
Digital transformation	436	78.5	0.08	0.81
Sustainability	414	282.2	0.01	0.82
Drug pipeline	190	26.8	0.00	0.87
Supply chain	135	15.5	0.00	0.81
ERP implementation	124	4.4	0.00	0.83
Energy transition	116	55.8	0.08	0.85
AI initiative	112	6.2	1.00	0.83
Data center	84	58.4	0.06	0.87
Automation	72	0.1	0.07	0.77
Cloud migration	56	71.1	0.09	0.85
Total	19,227	23,493.0	0.02	0.89

Sectoral structure. The composition of projects varies sharply across sectors. Figure 3 row-normalizes the project-type mix within each GICS sector. Utilities and Energy concentrate in plant construction and energy-transition projects; Information Technology and Communication Services in product launches and technology deployment; Health Care in R&D and drug pipelines; and Financials and Consumer sectors in M&A integration and geographic expansion. Table 3 reports the sector-level totals: every sector is well populated, with Industrials (2,708 projects), Health Care (2,475), and Information Technology (2,225) the most project-active.

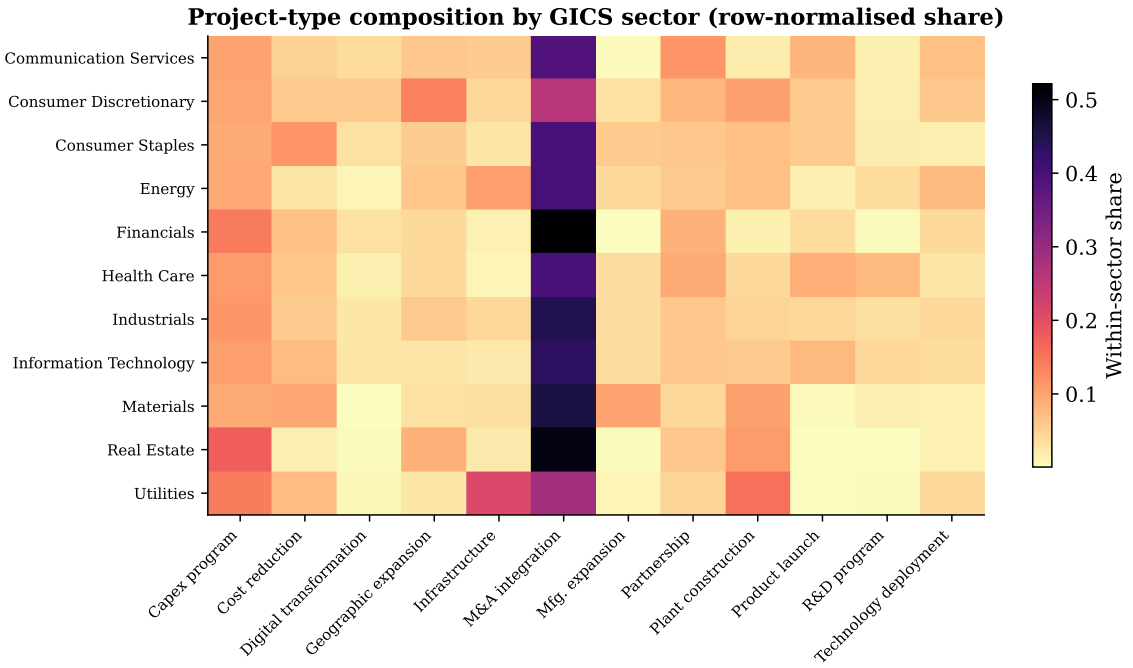
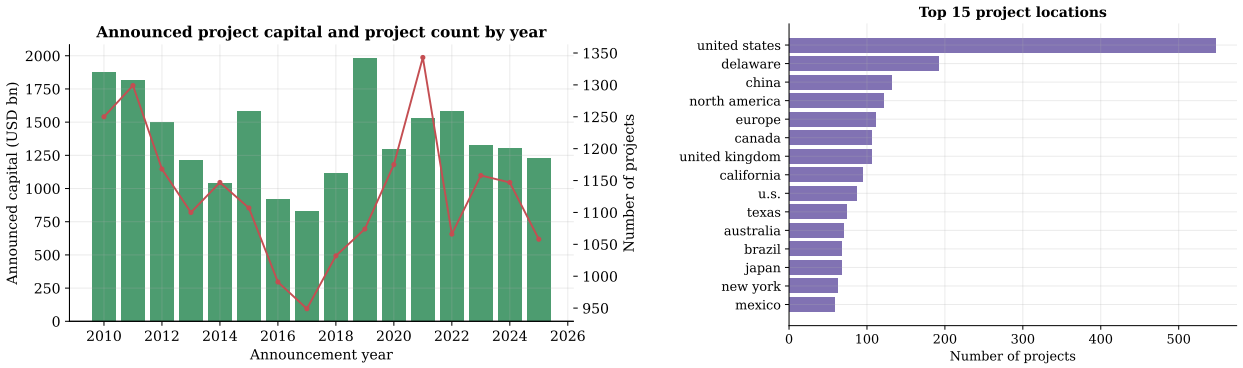


Figure 3: Project-type composition by GICS sector (row-normalized within-sector shares). Darker cells indicate a larger share of a sector's projects in that category.

Table 3: Sector breakdown: number of firms, projects, aggregate announced capital, and AI share.

GICS sector	Firms	Projects	Capital ($bn)	AI share
Industrials	79	2,708	2,128.6	0.03
Health Care	59	2,475	2,767.3	0.01
Information Technology	73	2,225	2,931.4	0.06
Consumer Discretionary	48	2,220	1,310.8	0.02
Utilities	31	1,956	2,725.9	0.01
Financials	76	1,908	4,832.3	0.01
Consumer Staples	36	1,483	1,332.5	0.01
Real Estate	31	1,400	901.5	0.00
Energy	21	1,109	1,267.8	0.01
Materials	26	914	643.5	0.00
Communication Services	20	829	2,651.5	0.03

Capital and geography. Figure 4a overlays announced capital and project counts by announcement year. Announced capital is lumpy and dominated by a handful of very large transactions in particular years, whereas the project *count* is smooth, underscoring why we model the extensive and intensive margins separately. Figure 4b shows the geographic concentration of projects, with the United States and a familiar set of global hubs accounting for most located projects.



(a) Announced capital and project count by year.

(b) Top 15 project locations.

Figure 4: Capital and geography of the corporate project economy.

4.1 Six stylized facts

The descriptive evidence can be distilled into six stylized facts that motivate the formal analysis and that, to our knowledge, have not previously been documented at this granularity:

1. **Project activity is voluminous and steadily growing.** Large public firms disclose on the order of 2,000–2,800 distinct project signals per year, growing at roughly 2.2% annually, with a level step-up after 2020.

2. **Acquisitions dominate the project mix.** M&A integration is the single largest category (38% of projects), reflecting both its economic weight and its disclosure salience among incumbents.

3. **Sectors specialize sharply.** Project composition is highly sector-specific — energy and utilities in physical builds, technology in deployments, health care in pipelines — with little cross-sector convergence in composition.

4. **Capital is concentrated in physical projects.** Disclosed amounts are heavily right-skewed and dominated by capital-intensive physical families; the median disclosed project is \$718m but the distribution spans four orders of magnitude.

5. **AI is frequent but small.** AI-oriented projects are proliferating rapidly in *count* yet remain individually modest in disclosed capital.

6. **Disclosure salience tracks scale.** The number of filings that mention a project rises monotonically with its announced amount, validating both the resolution procedure and the economic content of the extracted records.

5 Econometric Framework

The panel structure of the data supports a sequence of complementary models. We summarize the specifications here and report results in the following sections.

(i) Count models. Let y_{it} denote the number of projects firm i announces in year t. We model the conditional mean as

$$\mathbb{E}[y_{it} \mid x_{it}] = \exp(\alpha_i + \delta_t + x'_{it}\beta), \tag{1}$$

estimated by pseudo-Poisson maximum likelihood (PPML) with firm effects α_i and year effects δ_t. PPML is consistent under arbitrary distributional misspecification given a correctly specified mean, which makes it the natural workhorse for non-negative count outcomes. Because raw project counts are overdispersed, we also estimate a negative-binomial model and a pooled Poisson with sector and year effects. Standard errors are clustered by firm.

(ii) Hedonic capital regressions. For the 9,321 projects with a disclosed amount we estimate

$$\log(\text{amount}_j) = \gamma_{\text{type}(j)} + \theta_{\text{status}(j)} + \lambda\,\text{AI}_j + \mu' z_j + \alpha_{i(j)} + \delta_{t(j)} + \varepsilon_j, \tag{2}$$

where z_j collects project covariates (technology count, filing salience) and $\alpha_{i(j)}$, $\delta_{t(j)}$ are firm and year effects. The coefficients γ_{type} are interpreted as log-capital premia relative to a baseline category (M&A integration).

(iii) Diffusion. We model AI adoption at the firm-year level with a logistic regression for the probability of announcing at least one AI project, and we fit a logistic (S-curve) diffusion model $F(t) = K/(1 + e^{-r(t-t_0)})$ to the cumulative share of firms that have ever announced an AI project, recovering the saturation K, speed r, and inflection t_0.

(iv) Difference-in-differences. To test whether COVID-19 differentially accelerated digitalization, we define treatment as above-median *physical-project intensity* in the pre-period (2015–2019) and estimate a dynamic event study,

$$\text{DigitalShare}_{it} = \alpha_i + \delta_t + \sum_{\tau \neq 2019} \pi_\tau \left(\text{Treat}_i \times \mathbf{1}[t = \tau]\right) + u_{it}, \tag{3}$$

with 2019 omitted, alongside a canonical 2×2 specification. The leads π_τ for $\tau < 2019$ provide a parallel-trends test.

(v) Survival. We treat each multi-filing project as a spell that begins at first appearance and "fails" when it reaches completed status, right-censoring projects last seen as planned or in progress. We estimate Kaplan–Meier survival functions by project family and a Cox proportional-hazards model with project covariates.

(vi) Networks and representations. We analyze the degree distribution of the knowledge graph, detect communities in a technology co-occurrence network by greedy modularity maximization, and cluster the 384-dimensional project embeddings with K-means, selecting K by silhouette.

(vii) Prediction. Finally, we train a gradient-boosted tree classifier to predict "mega-projects" (announced amount above the sample median) from text-derived structural features, evaluated by 5-fold cross-validated ROC-AUC with permutation importance.

6 The Extensive Margin: Count Models

Table 4 reports the count-model estimates. Three results stand out.

First, project activity is **strongly persistent**. In the pooled Poisson the coefficient on the lagged project count is 0.100 ($p < 0.01$); within firms (column 2, firm and year effects) it remains positive and highly significant at 0.028, an incidence-rate ratio of about 1.03 per additional prior-year project. Persistence survives the inclusion of firm effects, indicating that it reflects within-firm dynamics — ongoing programs spawning follow-on initiatives — rather than fixed cross-firm heterogeneity alone.

Second, there is a **significant post-2020 increase** in announcement intensity. In specifications with sector (rather than firm) effects, the post-COVID indicator carries a coefficient of 0.208 ($p < 0.01$), implying roughly a 23% higher announcement rate after 2020, even as the secular year trend is mildly negative (-0.031 per year). The juxtaposition — a negative trend but a positive level shift after 2020 — suggests a structural break rather than a smooth acceleration.

Third, the data are **strongly overdispersed**. A Cameron–Trivedi auxiliary regression rejects equidispersion overwhelmingly ($t = 18.3$, $p < 10^{-70}$), and the negative-binomial dispersion parameter is $\hat{\alpha} = 0.50$. Overdispersion justifies the negative-binomial specification (column 3) and cautions against naive Poisson standard errors; our clustered, PPML-based inference is robust to it. Figure 5 visualizes the within-firm effects of persistence and accumulated AI experience.

Table 4: Panel count models of project-announcement intensity. Dependent variable: number of projects announced by firm i in year t. Cluster-robust standard errors (by firm) in parentheses. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

	(1) Pooled Poisson	(2) Poisson (firm+yr FE)	(3) Neg. Binomial
Projects$_{t-1}$	0.100***	0.028***	0.111***
	(0.004)	(0.005)	(0.005)
Cumulative AI$_{t-1}$	0.055***	-0.035	0.061***
	(0.012)	(0.023)	(0.013)
Post-COVID (2020+)	0.208***		0.209***
	(0.043)		(0.043)
Year trend	-0.031***		-0.031***
	(0.005)		(0.005)
Sector FE	Yes	–	Yes
Firm FE	–	Yes	–
Year FE	Yes	Yes	Yes
Observations	6,870	6,870	6,870
Log-likelihood	-14,751	-13,282	-13,533

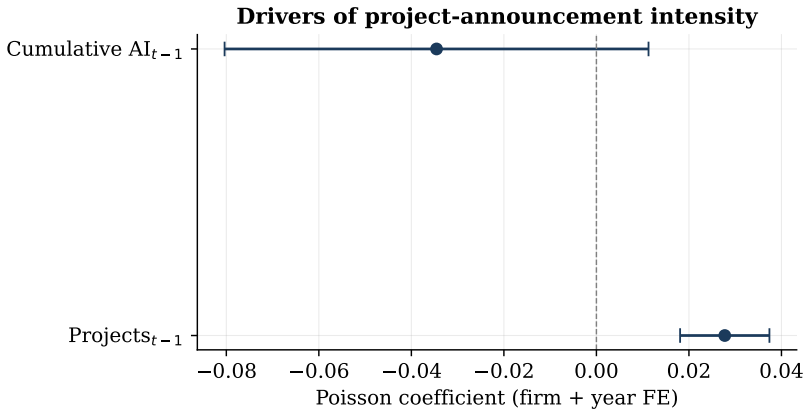


Figure 5: Within-firm drivers of project-announcement intensity (Poisson with firm and year effects, 95% confidence intervals).

The accumulated-AI-experience term is positive and significant in the pooled and negative-binomial models but attenuates to insignificance within firms, indicating that the cross-sectional association between AI experience and overall project activity is largely a firm-level (selection) phenomenon rather than a within-firm causal channel — AI-active firms are simply more project-active in general.

7 The Intensive Margin: What Drives Announced Capital?

Table 5 reports the hedonic log-capital regressions of equation (2). With firm and year effects (column 2) the model explains $R^2 = 0.318$ of the variation in log announced amount — substantial for project-level data — and the project-type fixed effects are jointly the dominant component.

Several findings are noteworthy. The number of distinct technologies attached to a project is positively associated with its size, consistent with larger projects being more technologically complex, though the within-firm estimate is imprecise. Filing salience (the log number of filings mentioning a project) is strongly positively associated with announced capital: larger projects are discussed more often, which both validates the resolution procedure and reflects genuine economic salience. Project *status* matters: relative to the omitted category, projects observed as in-progress or completed differ systematically in disclosed scale from those merely planned or mentioned.

Strikingly, AI projects are *not* larger than comparable projects; if anything the point estimate is negative (-0.18 in log points) and statistically insignificant ($p = 0.48$). This is an economically meaningful null: in this period AI initiatives are numerous and strategically prominent but individually modest in disclosed capital relative to plant builds, acquisitions, and infrastructure programs. Figure 6 plots the estimated capital premia by project type relative to M&A integration; capital-intensive families — infrastructure, plant construction, data centers, energy transition — carry the largest premia, while software, partnership, and cost-reduction projects sit below the baseline.

Table 5: Hedonic regressions for log announced capital (projects with a disclosed amount, $N = 9{,}321$). Baseline project type: M&A integration. Cluster-robust SE (by firm). $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

	(1) Sector+Year FE	(2) Firm+Year FE
AI project	-0.104	-0.184
	(0.306)	(0.261)
# technologies	0.033	0.065
	(0.050)	(0.044)
log(# filings)	0.728***	0.744***
	(0.049)	(0.043)
Status: in progress	0.303***	0.255***
	(0.078)	(0.064)
Status: completed		
Status: planned	0.507***	0.492***
	(0.082)	(0.073)
Project-type FE	Yes	Yes
Sector FE	Yes	–
Firm FE	–	Yes
Year FE	Yes	Yes
Observations	9,321	9,321
R^2	0.132	0.318

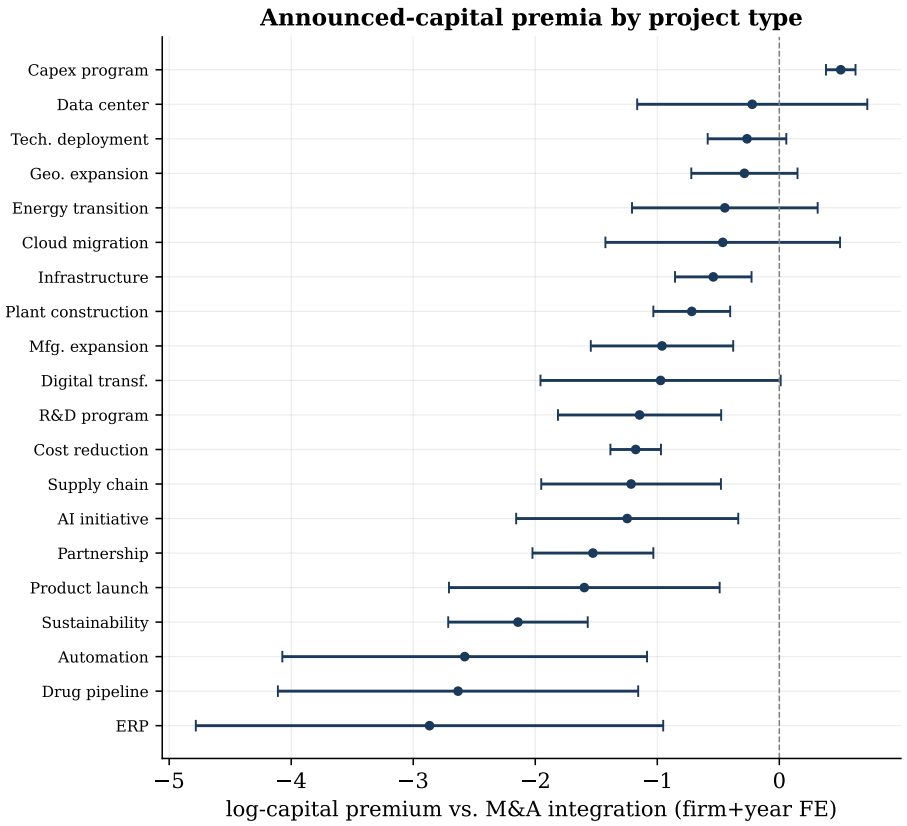


Figure 6: Announced-capital premia by project type relative to M&A integration (firm and year effects, 95% CI). Capital-intensive physical projects command the largest premia.

8 Technology Diffusion: The Rise of Corporate AI

Figure 7 summarizes the diffusion analysis. The left panel plots the cumulative share of firms that have ever announced an AI project together with the fitted logistic curve. The estimated saturation is $K = 0.42$, the speed is $r = 0.21$, and the inflection point is $t_0 = 2023.1$ — precisely when generative AI entered the corporate mainstream. The right panel shows the AI *share* of all new project announcements rising more than sixfold, from 0.7% in 2015 to 4.7% in 2025, with a clear acceleration after 2022.

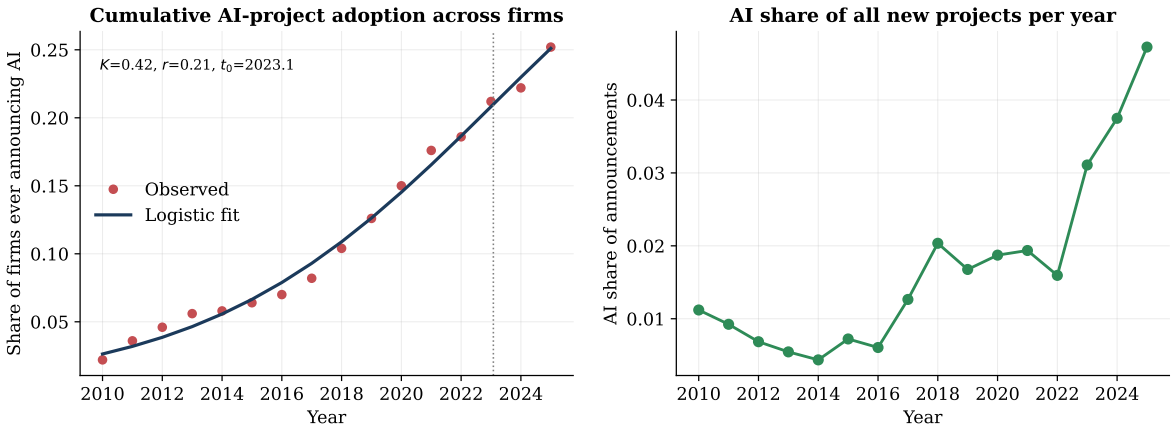


Figure 7: AI-project diffusion. Left: cumulative share of firms ever announcing an AI project, with a fitted logistic (S-curve) diffusion model; the vertical line marks the estimated inflection year (≈2023). Right: AI share of all new project announcements by year.

The firm-year logit (Table 6) confirms a positive secular trend in the probability of AI announcement (average marginal effect +0.25 percentage points per year) and a positive but imprecisely estimated post-COVID shift. Larger project portfolios are associated with higher AI-adoption probability, consistent with scale and absorptive capacity facilitating experimentation. That the diffusion is still on the steep part of its S-curve — the fitted saturation of 0.42 is far from unity — implies that corporate AI adoption, measured by disclosed projects, had substantial room to run at the end of the sample.

Table 6: Firm-year logit for the probability of announcing ≥ 1 AI project. Average marginal effects in the right column. Cluster-robust SE (by firm). $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

	Coefficient	Avg. marginal effect
Year trend	0.086***	0.0025
	(0.031)	(0.0009)
Post-COVID (2020+)	0.289	0.0084
	(0.273)	(0.0080)
Projects_{t-1}	0.118***	0.0035
	(0.022)	(0.0007)
Sector FE	Yes	
Observations	7,370	
Pseudo R^2	0.094	

9 COVID-19 and Digitalization: A Difference-in-Differences Test

A widespread narrative holds that the pandemic forced a rapid digital convergence, with physically intensive incumbents racing to catch up. We test this directly. Treatment is defined as above-median physical-project intensity in 2015–2019 (223 treated vs. 250 control firms); the outcome is the digital share of a firm-year's project announcements.

Figure 8 plots the dynamic treatment effects from equation (3). The pre-2019 leads are jointly indistinguishable from zero (F-test $p = 0.78$), supporting the parallel-trends assumption, and the post-2020 lags hover around zero. The canonical 2×2 estimate (Table 7) is -0.010 (s.e. 0.014, $p = 0.49$) — a precisely estimated null. We find *no evidence* that physically intensive incumbents differentially accelerated their digitalization after COVID-19 relative to already-digital firms. If anything, both groups digitalized in parallel. This tempers the convergence narrative: the pandemic may have raised digital activity broadly, but it did not measurably close the cross-firm gap in project composition.

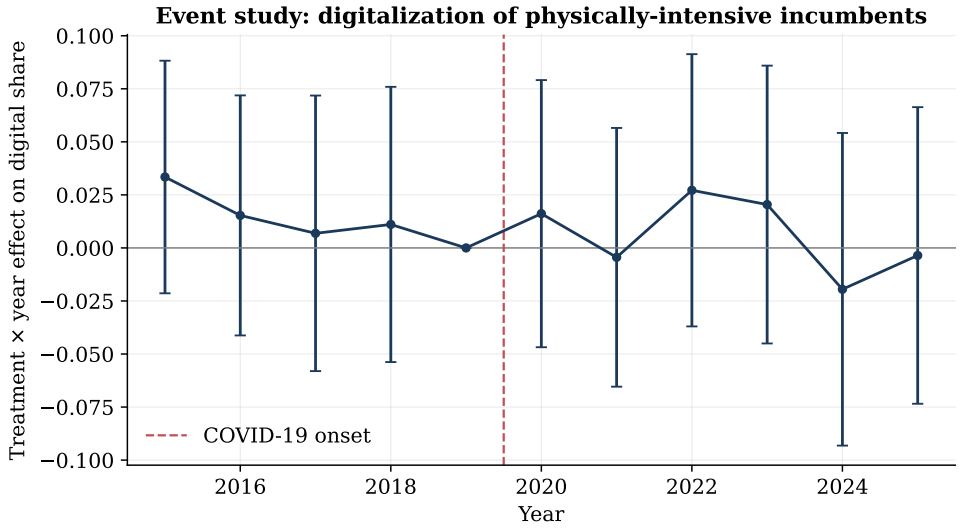


Figure 8: Event study of COVID-19 and digitalization. Points are treatment×year coefficients on the digital share of announcements relative to 2019 (95% CI). Flat pre-trends and post-period coefficients near zero indicate no differential digital acceleration of physically intensive incumbents.

Table 7: Difference-in-differences: COVID-19 and digital project share. Cluster-robust SE (by firm).

	Digital share
Treat × Post	-0.0100
	(0.0144)
Treat (pre-mean)	-0.0121
Year FE	Yes
Observations	3,845
R^2	0.016

10 Project Lifecycles: Survival Analysis

How long do corporate projects live? Among the 5,905 projects tracked across at least two filings (2,664 of which reach completed status), median lifecycle length varies dramatically by family. Figure 9 plots Kaplan–Meier survival curves — the probability that a project has *not* yet been completed as a function of months since its first appearance. M&A integrations resolve fastest (median

≈49 months), capital and construction projects next (≈84 months), while R&D / product programs (≈132 months) and especially digital / AI initiatives (≈144 months) persist far longer. The ordering is intuitive: acquisitions have a natural completion event, whereas digital transformation and AI are open-ended, continuously renewed programs that rarely declare victory.

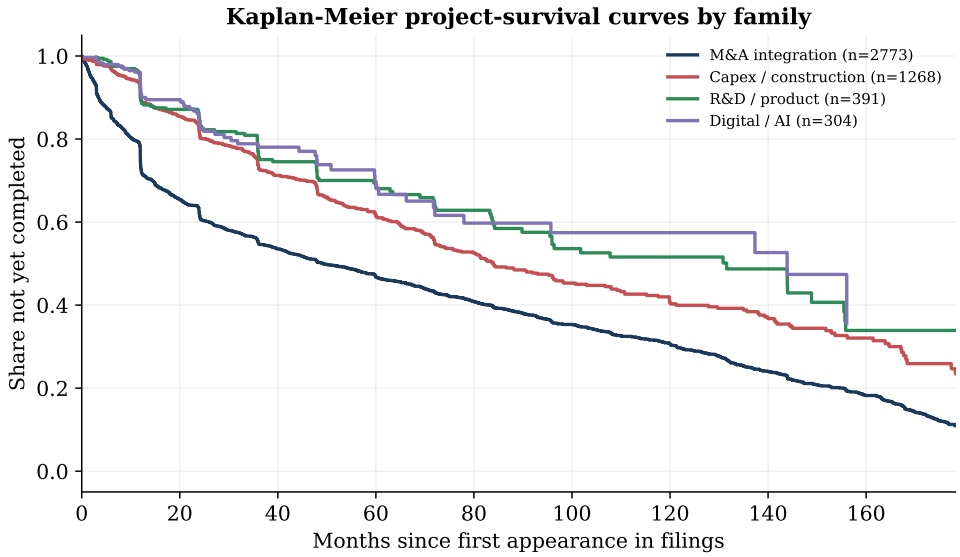


Figure 9: Kaplan–Meier survival curves by project family. M&A integrations complete fastest; digital and AI programs persist longest.

The Cox proportional-hazards model (Table 8) quantifies the covariate effects. Larger projects complete more slowly: a one-log-point increase in announced amount lowers the completion hazard by about 4% (hazard ratio 0.96, $p < 10^{-10}$), consistent with bigger initiatives being more complex and longer-lived. The AI indicator is not significantly associated with the completion hazard once project type and scale are controlled (HR 1.06, $p = 0.74$), reinforcing that the long lifecycles of AI programs operate through their *type* (open-ended digital programs) rather than an independent AI effect.

Table 8: Cox proportional-hazards model of time to project completion (hazard ratios). Project-type controls included. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Covariate	Hazard ratio	p-value
is ai	1.063	0.743
log amount	0.957***	0.000
n tech	0.850***	0.000
Project-type controls	Yes	
Projects (events)	2,664 of 5,905	

11 The Innovation Network

The knowledge graph offers a structural view of the corporate project economy. Figure 10 plots its degree distribution on log-log axes: the distribution is heavy-tailed with an estimated tail slope of -1.55, the signature of a scale-free-like structure in which a few hub nodes (large diversified firms, ubiquitous technologies) connect to a great many projects while most nodes have low degree.

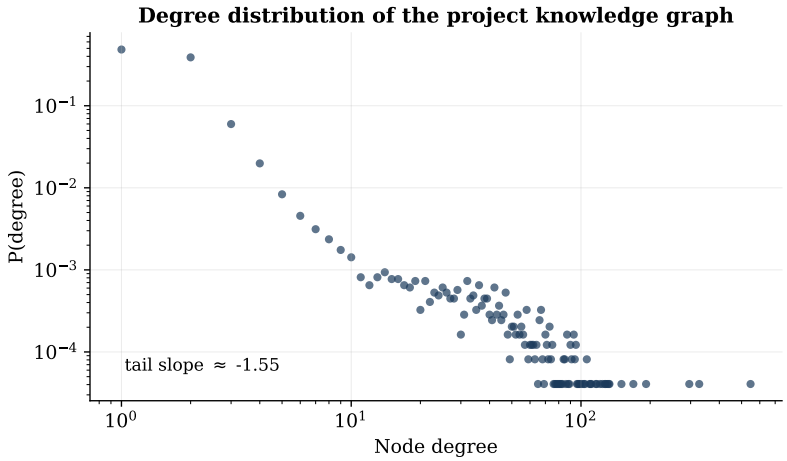


Figure 10: Degree distribution of the 32,226-node project knowledge graph (log-log). The heavy tail is consistent with a scale-free-like organization.

To uncover technology ecosystems we build a co-occurrence network in which two technologies are linked when they appear in the same project, restricted to the 60 most frequent technologies. Greedy modularity maximization recovers **four communities** with a high modularity of **0.52**, indicating well-separated ecosystems. Figure 11 visualizes the network; Table 9 lists the most central technologies. Strikingly, the highest eigenvector-centrality nodes are all *energy-transition* technologies — solar (0.57), renewable energy (0.49), wind (0.44), battery storage (0.26), natural gas, energy storage — which form the gravitational center of the corporate-innovation graph. Software/SaaS, cloud, AI, and semiconductors form a second, digital ecosystem; manufacturing/automation and connectivity form others. The centrality of the energy transition reflects both its capital intensity and its cross-sector reach: utilities, industrials, materials, and consumer firms all participate in it, whereas digital technologies, though pervasive, are more modular and less tightly co-bundled.

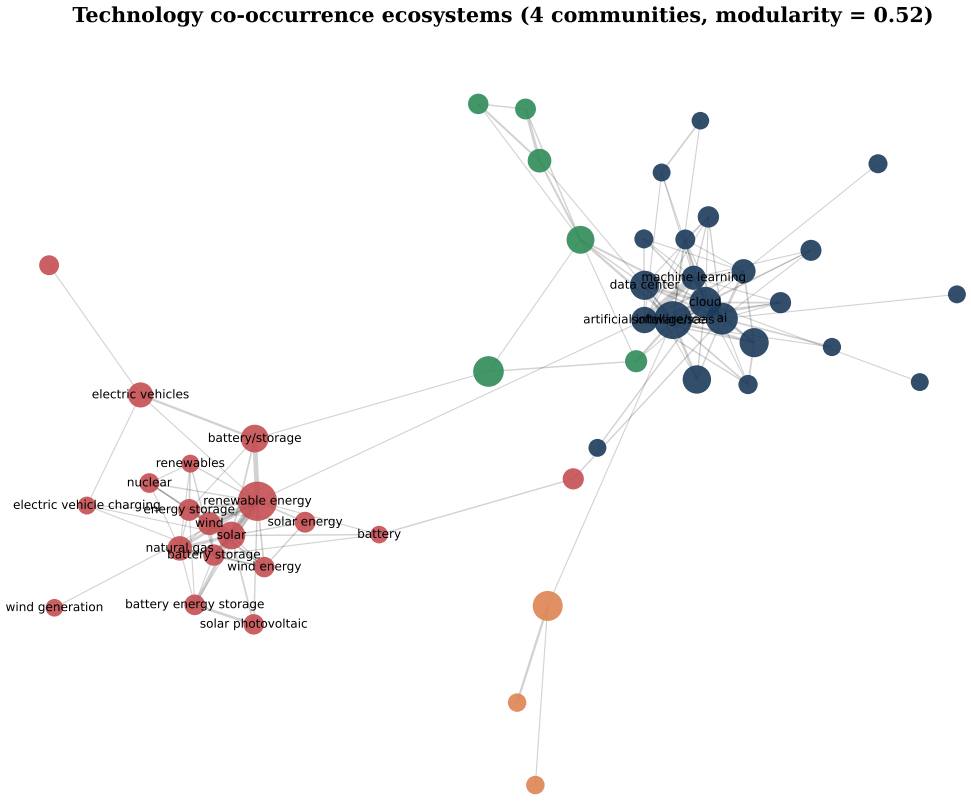


Figure 11: Technology co-occurrence network (top 60 technologies, edges for co-use in ≥ 2 projects). Node color denotes community; node size scales with project frequency. Four ecosystems emerge (modularity 0.52), with energy-transition technologies most central.

Table 9: Most central technologies in the co-occurrence network (eigenvector centrality) with project frequency and community membership.

Technology	Projects	Eigenvector cent.	Community
Solar	84	0.566	2
Renewable Energy	339	0.495	2
Wind	36	0.442	2
Battery Storage	24	0.264	2
Natural Gas	48	0.218	2
Energy Storage	27	0.199	2
Battery/Storage	85	0.158	2
Battery Energy Storage	20	0.119	2
Wind Energy	19	0.109	2
Nuclear	15	0.081	2
Solar Photovoltaic	19	0.074	2
Solar Energy	21	0.071	2
Renewables	8	0.059	2
Battery	7	0.052	2
Artificial Intelligence	67	0.037	1

12 Latent Themes: Embedding the Project Corpus

The formal taxonomy is imposed at extraction time. To recover *emergent* structure we cluster the 384-dimensional MiniLM project embeddings with K-means, selecting $K = 6$ by silhouette (Figure 13). Although the silhouette values are modest — expected for dense sentence embeddings of heterogeneous business text — the resulting clusters are interpretable and economically coherent (Table 10). They span an M&A/technology cluster, a health-care partnership and AI cluster, an energy/utilities construction cluster centered on renewables and storage, an industrial capex/ERP cluster, a consumer geographic-expansion cluster, and a consumer-staples cost-reduction and automation cluster.

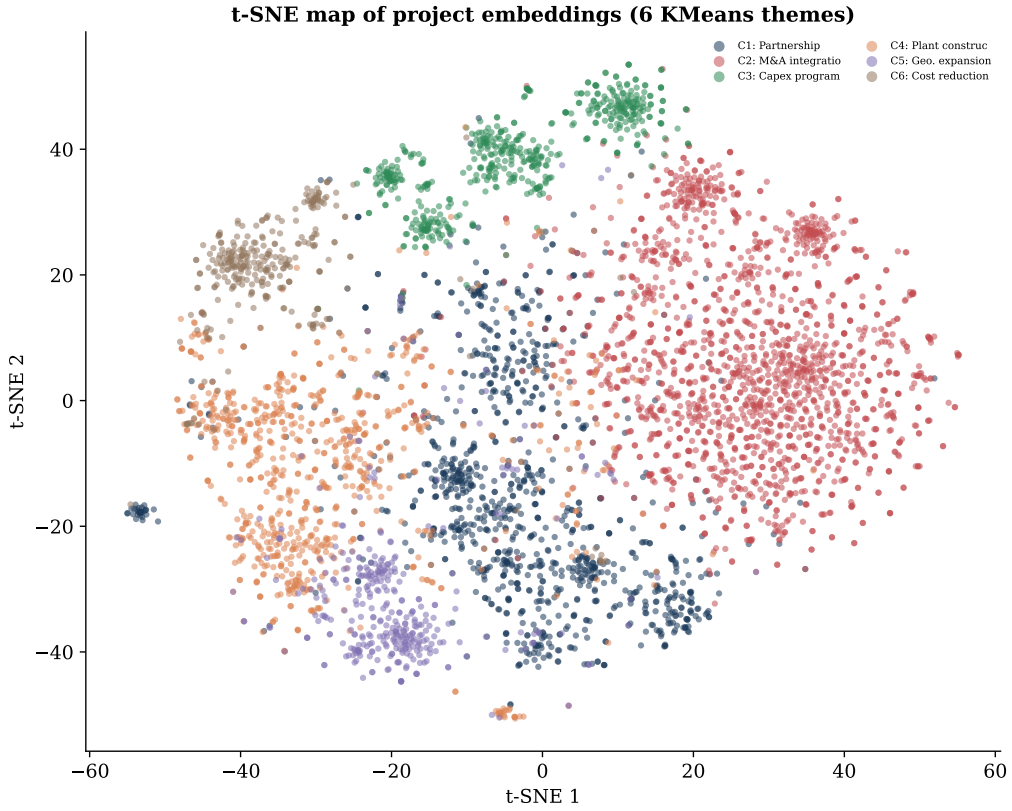


Figure 12: Two-dimensional t-SNE projection of project embeddings (stratified sample of 5,000 projects), colored by K-means cluster. Themes separate while overlapping at the boundaries, reflecting the continuous semantic space of corporate initiatives.

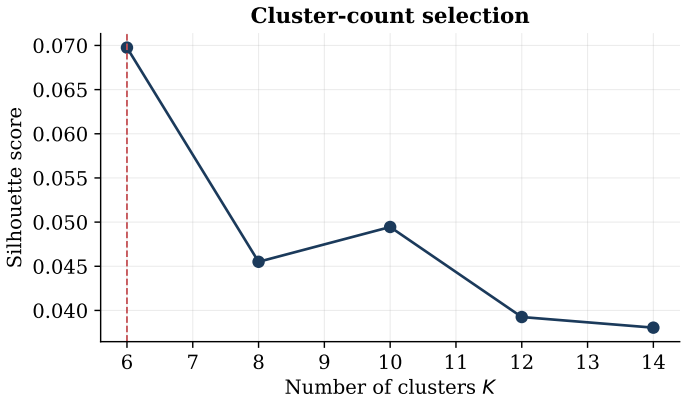


Figure 13: Silhouette score versus number of clusters; $K = 6$ is selected.

Table 10: Embedding clusters: size, dominant project type, top technologies, and aggregate announced capital. Clusters cross sector boundaries, capturing themes the taxonomy alone does not.

Cl.	Projects	Dominant type	Top technologies	Capital ($bn)
2	7,236	M&A integration	Software/Saas, Semiconductors, Cloud	14,840.0
1	4,093	Partnership	Software/Saas, Ai, Cloud	1,414.1
4	3,090	Plant construction	Renewable Energy, Solar, Battery/Storage	1,665.3
3	2,096	Capex program	Erp, Renewable Energy, Connectivity	4,559.5
5	1,551	Geo. expansion	Software/Saas, Data Center, Connectivity	212.0
6	1,161	Cost reduction	Renewable Energy, Semiconductors, Automati	802.2

Figure 12 shows the t-SNE projection of a sample of projects. The clusters separate visibly while overlapping at their margins, consistent with corporate initiatives occupying a continuous semantic space rather than discrete bins. That the unsupervised themes largely *recover* the supervised taxonomy — while recombining it across sectors — is an internal-validity check on the extraction: the embeddings and the LLM labels agree on the broad organization of the corpus.

13 Predicting Mega-Projects

Can the scale of a project be predicted from its qualitative description alone? We train a gradient-boosted tree classifier to distinguish "mega-projects" (announced amount above the $718m median) from smaller ones, using only text-derived structural features — project type, sector, status, technology count, filing and mention counts, AI flag, presence of a location, announcement year, and extraction confidence — with no leakage of the amount itself.

The model attains a 5-fold cross-validated **ROC-AUC of 0.743** and average precision of 0.722 (Figure 14, Table 11). Permutation importance ranks *project type* first, followed by *announcement year*, *sector*, and *filing count*. The predictive value of year and sector indicates that mega-projects cluster in particular periods and industries (e.g., recent data-center and energy build-outs), while the importance of filing count confirms that disclosure salience tracks economic scale. That qualitative

text features alone recover three-quarters of the discriminative information about project scale is itself evidence that the extracted attributes carry real economic content.

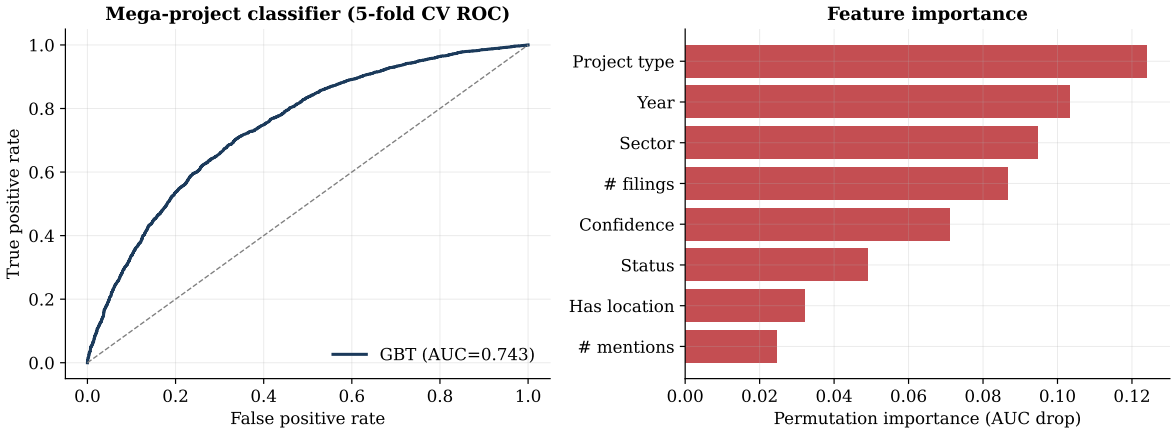


Figure 14: Mega-project classifier. Left: 5-fold cross-validated ROC curve (AUC = 0.74). Right: permutation feature importance (drop in AUC).

Table 11: Gradient-boosted mega-project classifier: feature-group importance and performance (N = 9,321 projects with disclosed amounts).

Feature group	Importance (AUC drop)
Project type	0.1240
Year	0.1033
Sector	0.0947
# filings	0.0867
Confidence	0.0711
Status	0.0491
Has location	0.0321
# mentions	0.0245
CV ROC-AUC	0.743
Avg. precision	0.722
Observations	9,321

14 Robustness and Validation

We summarize the checks that support the analysis.

Extraction quality. The mean extraction confidence is 0.78 and the median 0.90; records below 0.45 are dropped at the source. The agreement between the unsupervised embedding clusters and the supervised taxonomy (Section 12) provides an internal-consistency check: two independent representations of the same text organize it the same way. Spot auditing of high-amount records confirms that disclosed figures are grounded in the underlying text rather than hallucinated, a behavior the JSON-mode prompt explicitly enforces by forbidding unsupported numbers.

Estimator robustness. Because project counts are overdispersed, we report pooled Poisson, firm-fixed-effects PPML, and negative-binomial estimates side by side; the sign and significance of the persistence and post-2020 effects are stable across all three (Table 4). All panel models use cluster-robust standard errors by firm. The hedonic capital results are qualitatively identical under sector effects and under firm effects (Table 5), and the project-type ranking of premia is unchanged.

Parallel trends. The credibility of the difference-in-differences rests on parallel pre-trends, which we cannot reject (F-test $p = 0.78$); the dynamic event-study coefficients are flat before 2020 (Figure 8).

Level vs. composition. Throughout we privilege within-firm and relative comparisons over levels. The aggregate announced-capital figure is sensitive to a small number of very large transactions and to repeated mentions of multi-year programs; none of our regression conclusions depends on it, because the hedonic model uses log amounts with firm and year effects and the count models use the extensive margin.

Sample. The 500-firm corpus is weighted toward large, mature public companies, which is appropriate for studying the project economy of incumbents but limits external validity to small and private firms. Results should be read as describing large-cap U.S. public companies over 2010–2026.

15 Limitations

Several limitations warrant emphasis. *First*, our census measures *disclosure of* projects, not projects themselves; initiatives that are never discussed in filings are invisible, and disclosure incentives vary across firms and over time. *Second*, announced amounts are self-reported, inconsistently scoped (a figure may be total program cost, annual spend, or deal value), and frequently absent (only 48.5% of projects carry a number); we mitigate this with logs, fixed effects, and an explicit focus on relative magnitudes. *Third*, LLM extraction can err — both false positives (spurious projects) and false negatives (missed initiatives) — and although the confidence filter and grounding constraints reduce these, we do not have a hand-labeled gold standard large enough to estimate precision and recall precisely; constructing one is a priority for future work. *Fourth*, resolution merges mentions heuristically; aggressive merging understates project counts while conservative merging overstates them. *Fifth*, the survival analysis infers completion from disclosed status, which is itself a noisy signal. Finally, our identification in the difference-in-differences is quasi-experimental and the treatment is defined on pre-period behavior; we interpret the null cautiously, as the absence of a *differential* effect rather than the absence of any pandemic effect.

16 Discussion and Implications

The results have implications for research methodology, for the empirical study of corporate investment, and for practice.

For methodology. The paper is, in part, an existence proof: a single researcher, using a commodity LLM and open-source statistical tooling, can construct a population-scale, project-level panel from the unstructured regulatory record and subject it to the full apparatus of modern applied econometrics. The binding constraint in text-based empirical work has shifted from *extraction* to *validation and design*. Because extraction is now cheap and reproducible, the scarce inputs are a credible ground-truth benchmark and a research design that isolates economic mechanisms from disclosure artifacts. Our internal-consistency checks — the agreement of unsupervised embeddings with the supervised taxonomy, the monotone relation between disclosure salience and scale, and the stability of estimates across estimators — are a template, but they are not a substitute for a hand-labeled gold standard, which remains the most valuable complementary investment.

For the study of investment. Observing the project as a unit allows a clean separation of the extensive margin (how many initiatives a firm pursues) from the intensive margin (how large each is) and the compositional margin (which kinds). Our evidence suggests these margins are governed by different forces: the extensive margin is persistent and firm-driven with a macro level shift after 2020; the intensive margin is dominated by project *type* and physical capital intensity; and the compositional margin is strongly sectoral and, in the case of AI, diffusing along a classic S-curve. A natural agenda is to link these margins to financing conditions, governance, and ex-post outcomes — for instance, whether project-announcement intensity predicts future capital expenditure, productivity, or returns, and whether the market prices the project portfolio beyond aggregate guidance.

For the AI-diffusion debate. Our disclosure-based measure complements job-posting and patent-based measures and adds a distinct vantage point: the *strategic* framing of AI as a corporate project. The 2023 inflection and the still-low saturation imply that, as of the sample's end, corporate AI was early in its diffusion and that the count surge had not yet translated into large disclosed capital commitments — a pattern consistent with an experimentation phase preceding heavy build-out. Whether the subsequent data-center and compute investments convert AI from a high-count, low-capital category into a high-capital one is a question the same pipeline can answer as new filings arrive.

For practice. For analysts and policymakers, a structured project census enables monitoring that aggregate statistics cannot: tracking energy-transition build-outs across sectors in near real time, mapping partnership and supply-chain networks, or flagging firms whose project mix is shifting. The

knowledge graph and embeddings make the corpus queryable both structurally and semantically, turning a static disclosure archive into a living intelligence layer.

17 Conclusion

We have shown that the narrative sections of SEC filings, long treated as qualitative context, can be converted into structured, research-grade panel data on corporate projects at scale and at trivial marginal cost, and that the resulting data support a rich program of empirical analysis. From 153,894 filing sections we extracted 39,930 project mentions and resolved them into 19,227 projects across 500 firms, building a knowledge graph and semantic embeddings on top. The econometric analysis documents strong within-firm persistence of project activity and a significant post-2020 level shift; large capital premia for physically intensive project families and a surprising null for the scale of AI projects; an AI-diffusion path inflecting in 2023; a precisely estimated null for differential pandemic-driven digitalization; project lifecycles ranging from four years for acquisitions to twelve for digital programs; an energy-transition technology ecosystem at the center of the innovation network; and a text-only classifier that predicts project scale with an AUC of 0.74.

Beyond the specific findings, the broader message is that LLM-based information extraction reshapes what is measurable in empirical finance and the economics of innovation. The project — the actual unit of corporate decision-making — can now be observed at population scale. Natural next steps include constructing a human-labeled validation set to benchmark extraction precision and recall, extending the corpus to additional forms (S-1, DEF 14A, 20-F) and to the full universe of filers, linking projects to ex-post financial and market outcomes, and using the project panel to revisit classic questions about investment, financing constraints, and the returns to corporate innovation.

A Project Taxonomy and Variable Definitions

The extraction taxonomy comprises 21 mutually exclusive project types: manufacturing expansion, plant construction, data center, cloud migration, AI initiative, digital transformation, ERP implementation, R&D program, product launch, drug pipeline, M&A integration, geographic expansion, capex program, supply chain, automation, sustainability, energy transition, infrastructure, cost reduction, partnership, and technology deployment. Key engineered variables are defined as follows. **AI project**: a project whose type is *AI initiative* or whose technology list contains an AI term (AI, artificial intelligence, machine/deep learning, generative, LLM, neural). **Digital project**: an AI project or one whose type is among cloud migration, ERP, data center, automation, digital transformation, or technology deployment. **Announcement year**: the first calendar year in which the project appears in any filing. **Duration**: days between a project's first and last appearance. **Announced amount**: the resolved (aggregated) USD figure disclosed for the project, when present.

B Pipeline and Reproducibility

The extraction backend is OpenAI-compatible and was run with `gpt-4o-mini` at temperature 0 in JSON mode. The pipeline is incremental and resumable: mention identifiers are stable SHA-1 hashes of (accession, section, type, anchor), so re-runs use `INSERT OR REPLACE` semantics and never double-count. The full corpus run processed the 153,894 sections in 34 firm-batches with checkpointing; resolution, graph construction, embedding, and all analyses are deterministic given the resolved store. Analysis code is organized as ten numbered scripts (dataset construction; descriptives; count models; hedonic models; diffusion; difference-in-differences; survival; network; embeddings; prediction), each emitting figures, LaTeX table fragments, and a machine-readable results file. Estimation uses `statsmodels` (Poisson, negative binomial, logit, OLS, Cox PHReg), `scipy` (logistic curve fitting), `networkx` (community detection and centrality), and `scikit-learn` (K-means, t-SNE, gradient boosting, cross-validation).

References

Acemoglu, D., Autor, D., Hazell, J., & Restrepo, P. (2022). Artificial intelligence and jobs: Evidence from online vacancies. *Journal of Labor Economics*, 40(S1), S293–S340.

Babina, T., Fedyk, A., He, A., & Hodson, J. (2024). Artificial intelligence, firm growth, and product innovation. *Journal of Financial Economics*, 151, 103745.

Bass, F. M. (1969). A new product growth model for consumer durables. *Management Science*, 15(5), 215–227.

Bybee, L., Kelly, B., & Su, Y. (2023). Narrative-driven fluctuations in sentiment: Evidence from large language models. Working paper.

Crouzet, N., & Eberly, J. (2019). Understanding weak capital investment: The role of market concentration and intangibles. *NBER Working Paper* 25869.

Fazzari, S. M., Hubbard, R. G., & Petersen, B. C. (1988). Financing constraints and corporate investment. *Brookings Papers on Economic Activity*, 1988(1), 141–206.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3), 535–574.

Griliches, Z. (1957). Hybrid corn: An exploration in the economics of technological change. *Econometrica*, 25(4), 501–522.

Hassan, T. A., Hollander, S., van Lent, L., & Tahoun, A. (2019). Firm-level political risk: Measurement and effects. *Quarterly Journal of Economics*, 134(4), 2135–2202.

Hayashi, F. (1982). Tobin's marginal q and average q: A neoclassical interpretation. *Econometrica*, 50(1), 213–224.

Hidalgo, C. A., Klinger, B., Barabási, A.-L., & Hausmann, R. (2007). The product space conditions the development of nations. *Science*, 317(5837), 482–487.

Hoberg, G., & Phillips, G. (2016). Text-based network industries and endogenous product differentiation. *Journal of Political Economy*, 124(5), 1423–1465.

Korinek, A. (2023). Language models and cognitive automation for economic research. *NBER Working Paper* 30957.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics*, 45(2–3), 221–247.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, 66(1), 35–65.

Peters, R. H., & Taylor, L. A. (2017). Intangible capital and the investment-q relation. *Journal of Financial Economics*, 123(2), 251–272.